DOCUMENT RESUME

ED 380 786 CS 012 085

AUTHOR Stallman, Anne C.; And Others
TITLE Alternative Approaches to Vocabulary Assessment. Technical Report No. 607.
INSTITUTION Center for the Study of Reading, Urbana, IL.
PUB DATE Apr 95
NOTE 20p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Concurrent Validity; Intermediate Grades; Reading Research; Reliability; Standardized Tests; Test Validity; *Vocabulary Development
IDENTIFIERS *Alternative Assessment; *Word Knowledge

ABSTRACT

Interviews with children about their knowledge of a set of words was used to examine the concurrent validity of three paper-and-pencil measures of knowledge of these words--a standardized vocabulary test and two experimenter-designed tests. One experimenter-designed test, the Levels test, had three multiple-choice items per word that targeted three different levels of word knowledge. The other was a forced-choice contexts test with five items per word, each requiring a decision about whether the word was used appropriately in the context. Subjects were 50 students from two heterogeneously grouped fifth-grade classrooms in a midwestern school district. All three paper-and-pencil measures showed acceptable levels of reliability. When subjects were used as the unit of analysis, the interview was more highly correlated with the standardized test and the Levels test than with the Contexts test. When the word was used as the unit of analysis, the interview correlated more highly with the Contexts and the Levels test than with the standardized test. These results are interpreted as indicating that standardized measures are more effective at discriminating among students upon the basis of their overall ability, but less accurate as measures of how much the students know about particular words. The Contexts test has the advantages of the highest reliability of the three measures, as well as the greatest instructional validity. (Contains 25 references and 7 tables of data.) (Author/RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 380 786

Technical Report No. 607

ALTERNATIVE APPROACHES TO VOCABULARY ASSESSMENT

Anne C. Stallman
P. David Pearson
William E. Nagy
Richard C. Anderson
Georgia Earnest García

University of Illinois at Urbana-Champaign

April 1995

Center for the Study of Reading

TECHNICAL REPORTS

College of Education
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

2501208S

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

F. Lehr

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

CENTER FOR THE STUDY OF READING

Technical Report No. 607

ALTERNATIVE APPROACHES TO VOCABULARY ASSESSMENT

Anne C. Stallman
P. David Pearson
William E. Nagy
Richard C. Anderson
Georgia Earnest García

University of Illinois at Urbana-Champaign

April 1995

College of Education
University of Illinois at Urbana-Champaign
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

MANAGING EDITOR
Technical Reports
Fran Lehr

MANUSCRIPT PRODUCTION ASSISTANT
Delores Plowman

Abstract

Interviews with children about their knowledge of a set of words was used to examine the concurrent validity of three paper-and-pencil measures of knowledge of these words--a standardized vocabulary test and two experimenter-designed tests. One experimenter-designed test, the Levels test, had three multiple-choice items per word that targeted three different levels of word knowledge. The other was a forced-choice contexts test with five items per word, each requiring a decision about whether the word was used appropriately in the context.

All three paper-and-pencil measures showed acceptable levels of reliability. When subjects were used as the unit of analysis, the interview was more highly correlated with the standardized test (r = .75) and the Levels test (r = .76) than with the Contexts test (r = .68). When the word was used as the unit of analysis, the interview correlated more highly with the Contexts test (r = .70) and the Levels test (r = .70) than with the standardized test (r = .55). These results are interpreted as indicating that standardized measures are more effective at discriminating among students on the basis of their overall ability, but less accurate as measures of how much the students know about particular words. The Contexts test has the advantages of the highest reliability of the three measures, as well as the greatest instructional validity.

ALTERNATIVE APPROACHES TO VOCABULARY ASSESSMENT

Because of the importance ascribed to test scores in the United States, educators often feel compelled to instruct children using materials and methods that mirror those used on the tests (Farr & Carey, 1986; Pearson & Valencia, 1987; Valencia & Pearson, 1987, 1988). As a result, constructing tests that are reflective of sound research and instructional practices has become an important issue in education. Although there have been attempts in the past few years to update assessments of reading comprehension to align them more closely with findings from exemplary reading research and practice, similar attempts have not been made with assessments of vocabulary. This brings into question the notion of the construct validity of current vocabulary assessments and raises the question: Do they test word knowledge?

The purpose of this study was to examine the concurrent validity of three paper-and-pencil measures of vocabulary knowledge to determine which measure most closely reflected students' actual word knowledge.

Background

Most reading tests given to students include a section designed to measure vocabulary knowledge. The inclusion of such a section reflects the long-standing research tradition that has documented the strong correlational relationship between vocabulary knowledge and reading comprehension (Anderson & Freebody, 1981; Dale & O'Rourke, 1986; Davis, 1944, 1968; Spearritt, 1972; Thorndike, 1973; Thurstone, 1946). Studies within this tradition have demonstrated consistently that word knowledge is strongly related to reading comprehension; in fact, these studies show that word knowledge is the single best predictor of an individual's ability to comprehend a text (Johnston, 1984; Johnston & Pearson, 1982). The inclusion of measures of vocabulary knowledge in reading assessments, therefore, seems to be a logical step.

Most of the current assessments of vocabulary knowledge, however, focus on students' knowledge of word definitions. They present words in isolation, and assume that students' word knowledge is captured by their ability to identify a synonym. Although some tests have attempted to address issues relating to the importance of context in vocabulary knowledge, the *contexts* that are included on the tests do not require students to use the additional information to answer the questions. For example, on some current standardized tests, students are instructed to choose the word that means the same or about the same as the underlined word. For example:

whole story
A. true
B. short
C. unusual
D. complete

and

Elevation is the same as _____.
A. climate
B. altitude
C. region
D. direction

Research has shown that it is possible for students to get the right answer on items such as these by using only partial knowledge of the words (Curtis, 1987). Given that it is *in-depth knowledge* of the concepts represented by the words that has been shown to be a critical factor in reading comprehension (Anderson, 1985; Anderson & Freebody, 1981; Anderson & Pearson, 1984, 1988; Beck, Perfetti, & McKeown, 1982; Johnson & Pearson, 1984; Mezynski, 1983; Nagy, 1988; Pearson, 1986), the value of this type of vocabulary testing is brought into question. Furthermore, these traditional types of assessment items do not discriminate between individuals with partial knowledge and those with more complete knowledge.

Because there is some evidence that vocabulary acquisition is incremental in nature, it seems important to have measures of vocabulary knowledge that are sensitive to different levels of word knowledge. Although students may not demonstrate complete knowledge of a word's meaning after encountering it, either during a natural reading situation or through an instructional intervention, they should be given credit for smaller gains in their knowledge of the word's meaning. There have been attempts to address the issue of incremental gains in word knowledge. Nagy, Herman, & Anderson (1985) developed a multiple-choice test that required increasing levels of sophistication of word knowledge. These items varied in the degree to which the distractors were related to the correct answer, but the correct answer was always the same.

In this study, we examined the concurrent validity of three paper-and-pencil measures of vocabulary knowledge to determine which measure most closely reflected students' actual word knowledge. One test consisted of traditional standardized test items and the other two were designed to capture students' levels of knowledge about individual words. Test scores were compared to a criterion measure, which was based upon information students provided about the words during an interview. In essence, we evaluated how well each measure mirrored the interview. The logic of this line of inquiry is that if efficiency were not an issue, interviews would yield the richest data about individuals' conceptual knowledge for any given domain. However, because there is a need for at least some assessment efficiency, an important issue is which testing format, among a wide array of competing formats, is the best surrogate for the interview approach (Anderson & Freebody, 1983).

Method

Subjects

Fifty students from two heterogeneously grouped fifth-grade classrooms in a midwestern school district participated in the study. Demographic data for the district and school are presented in Table 1.

[Insert Table 1 about here.]

Materials

A stratified sample of 25 words was selected from Levels E through J of the Comprehensive Tests of Basic Skills (CTBS) (1981). Five words were chosen randomly from each level to represent a wide range of difficulty.

The children were interviewed about their knowledge of the 25 words and took three paper-and-pencil tests over them. On one test, the items were presented as on the CTBS (Standard Test). This represents the types of items that are traditionally used on norm-referenced tests. Students were told to choose the word that means the same or about the same as the underlined word. For example:

infinite choices
toward
countless
unrealistic
independent

In another format, each word was tested at three different levels of knowledge sophistication (Levels Test). Level I items required minimal knowledge. The student was asked to choose the pair of words that the target word went with. For example:

flee goes with:

picture	gentle	go	don't know
photo	silky	leave	

Level II items required some general knowledge of the meaning of the word. For example:

flee means
a. walk beside
b. run from
c. carry gently
d. tiptoe quietly
e. don't know

Level III items required precise knowledge of the meaning of the word. For example:

flee means
a. sign out
b. hide under
c. escape quickly
d. leave quietly
e. don't know

In the third format, the students were asked to respond *Yes*, *No*, or *Don't Know* to questions in which the target word was used (Contexts Test). Five questions for each word were presented in a random order. The questions required different levels of word knowledge. For example:

Do toss like to fish?	Yes	No	Don't Know
Can a bell toss?	Yes	No	Don't Know
Can a person toss a real house?	Yes	No	Don't Know
Is tossing a way of throwing?	Yes	No	Don't Know
Is tossing something you do gently?	Yes	No	Don't Know

Procedures

The students were interviewed individually about their knowledge of the words. They were shown a word on a card and asked to read it. Pronunciation was corrected when necessary. Then the students were asked what the word meant. They were prompted to give additional information and to use the word in a sentence. The students were encouraged to give any information they could think of, even if they were not sure it was complete.

The children's responses were tape recorded and notes were taken during the interviews. Results showed that correcting pronunciation of the words was helpful in allowing students to get the meaning

of a word only 4% of the time. In other words, if a word was not in the student's reading vocabulary, it was not likely to be in his or her oral vocabulary either.

A week after the interviews were completed, the students took the three paper-and-pencil tests over the same words they had been asked about in the interviews. The students completed the Levels Test on the first day of testing, the Contexts Test on the second day, and the Standard Test on the third day. The tests were administered to all 50 children as a group. Four forms of each test, in which the items were presented in different random orders, were used to eliminate order effects of the items and to reduce the possibility of copying. The students were allowed to take as much time as they needed to complete the tests. For the slowest students, the Standard Test took 15 minutes, the Levels Test took 18 minutes, and the Contexts Test took 21 minutes.

Scoring

The responses students gave during the interview were scored using the following scale:

0 No correct information was given.
1 Some correct information was given; the student had a vague idea of some aspect of the word's meaning.
2 Mostly correct information was given but some aspect was missing.
3 The word's complete meaning was given.

The interrater reliability for scoring the interview responses was 97% for three independent raters.

Paper-and-pencil tests. All three paper-and-pencil tests were scored in the same manner. Students received 1 for choosing a keyed response, 0 for choosing *Don't Know*, and -1 for choosing a response that was not keyed. The means, standard deviations, and ranges are displayed in Table 2. In the top part of the table, scores represent raw scores. To allow for more direct comparisons among the measures, the scores in the bottom part of the table have been converted to percentage correct. As can be seen from mean percentage correct scores in the table, the two multiple-choice tests, the Standard Test and the Levels Test, were somewhat more difficult for the students than was the Contexts Test. There were no ceiling or floor effects on any of the measures, however.

[Insert Table 2 about here.]

In addition, students were asked to make comments about what they thought about the different test formats. Most students said the Levels Test was "easy in some parts and hard in others" and that the Contexts Test was "long, but easy." They felt that the Standard Test was unfair because there was no *Don't Know* option, and they could not understand why they should guess if they did not know the word being tested.

Results and Discussion

This study was designed to evaluate how closely each of the paper-and- pencil measures mirrored the information students gave in an interview. It was assumed that the interview represented the most complete picture of the student's knowledge of the words; therefore, the interview was used as the criterion measure. There are several questions of interest:

1. How reliable is each measure?
2. How well does each paper-and-pencil measure correlate with the interview?
3. How much of the variance in the interview scores can be accounted for by the paper and pencil measures?

BEST COPY AVAILABLE

The data from this study can be approached in two different ways: one using subject as the unit of analysis and the other using word as the unit of analysis. These data were examined both ways.

Reliability

Cronbach's alpha was used to compute reliability coefficients for each of the paper-and-pencil measures (see Table 3). The reliabilities were computed using one score for each word. While the reliabilities of the individual levels in the Levels Test are lower than those of the other tests, the reliability of the total Levels Test is higher than the reliabilities of the subtests, which may argue for using the test as a whole rather than looking at performance on the levels separately. Given a criterion of .80 or higher as minimally acceptable, all three measures are reliable.

[Insert Table 3 about here.]

Concurrent Validity

Data from this study were analyzed in two ways, first using the subject as the unit of analysis, and second, using the word as the unit of analysis. When the subject is used as the unit of analysis, each subject is assigned a single score computed by summing that subject's score for each word. In this type of analysis, a correlation between two tests is a measure of how well they agree about the relative overall performance of individuals. When the word is used as the unit of analysis, each word is assigned a single score computed by summing the scores of each subject for that word. In this type of analysis, a correlation between two tests is a measure of how well they agree about the relative overall difficulty of the words.

Subject as the unit of analysis. When subject is used as the unit of analysis, it is possible to examine the relationships among the measures in terms of how they discriminate among individuals. The scores used in these analyses are computed by summing scores across words for each individual. Correlations were computed to examine the relationships among the measures (see Table 4). As would be expected for correlations among measures of verbal performance, the correlations among all the measures are moderate to strong, ranging from .67 to .85. The Standard Test and the Levels Test are correlated equally with the interview, while the correlation between the Contexts Test and the interview is somewhat lower. This result can be interpreted to mean that the Standard Test and the Levels Test are somewhat better than the Contexts Test at discriminating among individuals. The strong correlation between the Levels Test and the Contexts Test indicates that the information gained from these measures is similar, and the lower correlations between the Standard and the Contexts Test and the Levels Test indicate that somewhat different information is being tapped by the Standard Test.

[Insert Table 4 about here.]

A series of regression analyses was conducted to examine how much of the variance in the interview scores could be accounted for by the paper-and-pencil measures, and the extent to which the measures accounted for unique variance. Four separate analyses were run in which the dependent measure was the interview and the paper-and-pencil measures were entered in different orders as the predictors (see Table 5). The three measures together accounted for 66% of the variance in the interview scores, and the amount of variance accounted for by each of the paper-and- pencil measures individually was statistically significant irrespective of the order of entry. When it was entered first, the Standard Test scores accounted for more than half of the variance in the interview scores. When it was entered last, it still accounted for more than four times as much unique variance as either the Levels Test or the Contexts Test did when they were entered in the last position. When subject was the unit of analysis, the Standard Test accounted for more unique variance in the interview scores than either the Levels Test or the Contexts Test.

[Insert Table 5 about here.]

Word as the unit of analysis. The same correlations and regression analyses were performed using word as the unit of analysis so as to examine the relationships among the measures in terms of how they discriminate among knowledge of individual words. The scores used in these analyses were computed by summing scores for words across individuals. These analyses paint a very different picture from the analyses using subject as the unit of analysis. In these analyses, the correlations among the interview, the Contexts Test and the Levels Test are much higher than the correlations between the Standard Test and the other measures (see Table 6). These correlations indicate that the Contexts Test and the Levels Test are better than the Standard Test at discriminating among subjects' knowledge of words. The consistently strong correlations among the Contexts Test, the Levels Test, and the interview indicate that they are all tapping similar information while the lower correlations with the Standard Test indicates that it is tapping somewhat different information.

[Insert Table 6 about here.]

The Contexts Test and the Levels Test are more valid indicators of actual word knowledge than the Standard Test because of the properties of the tests. Items were chosen for inclusion on the Contexts and Levels Tests because they represented particular knowledge of the word's meaning: Students who knew the particular aspect of the word's meaning targeted by an item would get it right and students who did not know would get it wrong. However, items on the Standard Test are chosen based on the psychometric properties of the item. Items included on this type of test are chosen because they discriminate among individuals, not necessarily because of the difficulty of the word being tested. An easy word may be included because the distractors make it difficult and a difficult word may be included because the distractors give away the correct answer. Therefore, while standardized tests may discriminate among individuals, they are not as useful as measures of specific word knowledge (Anderson & Freebody, 1983).

The same four regressions were run using word as the unit of analysis as were run using subject as the unit of analysis. In these analyses, the dependent measure was the interview and the paper-and-pencil measures were entered in different orders as predictor variables (see Table 7). The three measures together accounted for 61% of the variance in the interview scores, and the amount of variance accounted for by each of the paper-and-pencil measures individually was statistically significant irrespective of the order of entry. However, both the Contexts Test and the Levels Test consistently accounted for about twice as much of the variance in the interview scores as did the Standard Test.

[Insert Table 7 about here.]

Conclusion

Even though an interview format probably provides the richest picture of students' knowledge of the target vocabulary, it is time consuming to administer and difficult to score reliably (Anderson & Freebody, 1983). The purpose of this study was to evaluate how well each of three paper- and-pencil measures mirrored information gathered in the interviews.

The Standard Test is a traditional, widely used measure. It had an acceptable reliability and a strong relationship to the interview when subject was used as the unit of analysis but a weaker relationship when word was the unit of analysis. This is not surprising, because the purpose of tests like the Standard Test is to discriminate among individuals, and the tests have been carefully designed to achieve that purpose. However, the Standard Test is less effective for measuring knowledge of individual words.

The Levels Test had a high reliability and showed a fairly strong relationship to the interview in both analyses. Based on these analyses, the Levels Test would be a reasonably good choice for use in assessing levels of word knowledge.

The Contexts Test had the highest reliability of the three measures. In addition, its relationship to the interview, using word as the unit of analysis, was equal to that of the Levels Test. Therefore, it would also be a good choice for assessing levels of word knowledge. However, the Contexts test does have an advantage over the Levels Test. The Levels Test looks like traditional multiple choice vocabulary tests, but the Contexts Test uses the target words in contexts, which is more closely in line with recommended vocabulary instructional techniques. This is relevant because of the strong link between assessment and instruction. The instructional validity of the Contexts Test is important because "teaching to" this type of test would result in instructional activities that focus on the integration of information about words across a variety of contexts. Other types of tests have the potential of encouraging unproductive instructional activities. One way to encourage change in instructional techniques is to change the types of assessments that are being used, and the Contexts Test has that potential.

The success of the Contexts Test in this study leads to speculation about its potential for use in future research and in classrooms. A test that can measure different levels of word knowledge successfully could be useful in different types of research. It could be used to evaluate the effectiveness of instructional interventions as well as to document the types or levels of vocabulary knowledge to which different activities contribute. The Contexts Test also shows promise as a tool for classroom use. Given the fact that assessments communicate something about the nature and what is important about the construct being assessed (in this case, vocabulary knowledge), as well as the strength of the link between assessment and instruction, it is important that the assessments reflect current thinking and research about the relationship between vocabulary knowledge and reading comprehension and how vocabulary knowledge is acquired. Current vocabulary research emphasizes the importance of contextualizing vocabulary instruction and the importance incremental knowledge acquired incidentally. The Contexts Test has the advantages of assessing vocabulary knowledge in contexts and of being sensitive to partial knowledge of words.

References

Anderson, R. C. (1985). Role of reader's schema in comprehension, learning, and memory. In H. Singer & R. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 372-384). Newark, DE: International Reading Association.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutton (Ed.), *Advances in reading/language research: A research annual* (pp. 231-256). Greenwich, CT: JAI Press.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-291). New York: Longman.

Anderson, R. C., & Pearson, P. D. (1988). A schema-theoretic view of basic processes in reading comprehension. In P. L. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 37-55). Cambridge: Cambridge University Press.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*, 506-521.

Comprehensive Tests of Basic Skills (Form U). (1981). Monterey, CA: CTB/McGraw-Hill.

Curtis, M. E. (1987). Vocabulary testing and vocabulary instruction. In M. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 37-52). Hillsdale, NJ: Erlbaum.

Dale, E., & O'Rourke, J. (1986). *Vocabulary building: A process approach*. Columbus, OH: Zaner-Bloser.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika, 9*, 185-197.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499-544.

Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* (2nd ed.). Newark, DE: International Reading Association.

Johnson, D., & Pearson, P. D. (1984). *Teaching reading vocabulary* (2nd ed.). New York: Holt, Rinehart & Winston.

Johnston, P. (1984). Background knowledge and reading comprehension test bias. *Reading Research Quarterly, 19*, 219-239.

Johnston, P., & Pearson, P. D. (1982). *Prior knowledge connectivity and the assessment of reading comprehension* (Tech. Rep. No. 245). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53*, 253-279.

Nagy, W. (1988). *Teaching vocabulary to improve reading comprehension*. Newark, DE: International Reading Association.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233-253.

Pearson, P. D. (1986). Twenty years of research in reading comprehension. In T. E. Raphael (Ed.), *The context of school-based literacy* (pp. 43-62). New York: Random House.

Pearson, P. D., & Valencia, S. W. (1987). Assessment, accountability, and professional prerogative. In J. Readance & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives* (pp.3-16). Rochester, NY: The National Reading Conference.

Spearritt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8*, 92-111.

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries*. New York: Wiley.

Thurstone, L. L. (1946). A note on a reanalysis of Davis' reading tests. *Psychometrika, 11*, 185-188.

Valencia, S. W., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher, 40*, 726-732.

Valencia, S. W., & Pearson, P. D. (1988). Principles for classroom comprehension assessment. *Remedial and Special Education, 9*, 26-35.

Table 1

School and District Demographic Information

	School	District
Ethnicity		
White	77.2%	66.9%
Black	20.6%	27.6%
Hispanic	2.2%	3.5%
Total Enrollment	325	7075
Low-Income	47.7%	39.9%
Limited English Proficiency	1.2%	2.6%
Attendance Rate	94.9%	92.8%
Student Mobility	39.0%	27.5%
Non-Promotion Rate	1.8%	2.9%

Table 2

Descriptive Statistics for Each Measure

Test	Mean	Standard Deviation	Range
		Raw Scores	
Interview	36.88	8.02	20.00-56.00
Standard	8.73	7.87	-17.00-22.00
Contexts	55.70	20.50	3.00-90.00
Levels	26.94	15.05	-6.00-56.00
		Percentage Correct	
Interview	49.39	10.72	26.67-74.67
Standard	67.17	16.11	12.00-92.00
Contexts	57.97	14.09	28.00-82.40
Levels	53.92	14.81	17.33-82.67

Table 3

Cronbach's Alpha for Each Paper and Pencil Measure

Test	Alpha
Levels Test	
Total	.84
Level 1	.80
Level 2	.79
Level 3	.75
Contexts Test	.89
Standard	.81

Table 4

Correlations Among Measures: Subject as the Unit of Analysis

	Standard	Contexts	Levels	Level 1	Level 2	Level 3
Interview	.75	.68	.76	.67	.78	.72
Standard		.60	.74	.70	.73	.70
Contexts			.85	.84	.80	.81
Levels				.96	.96	.85
Level 1					.86	.93
Level 2						.88

Table 5

Variance in Interview Accounted for by Paper and Pencil Tests: Subject as the Unit of Analysis

Order of Predictors	Multiple R^2	R^2 Change
Standard	.56	.56*
Contexts Test	.64	.08*
Levels Test	.66	.02*
Standard	.56	.56*
Levels Test	.65	.09*
Contexts Test	.66	.01*
Levels Test	.57	.57*
Contexts Test	.58	.01*
Standard	.66	.08*
Contexts Test	.46	.46*
Levels Test	.58	.12*
Standard	.66	.08*

*$p < .05$.

Table 6

Correlations Among Measures: Word as the Unit of Analysis

	Standard	Contexts	Levels	Level 1	Level 2	Level 3
Interview	.55	.70	.70	.62	.64	.56
Standard		.50	.55	.47	.46	.47
Contexts			.72	.63	.65	.55
Levels				.85	.87	.82
Levels					.65	.51
Levels						.55

Table 7

Variance in Interview Accounted for by Paper and Pencil Tests: Word as the Unit of Analysis

Order of Predictors	Multiple R^2	R^2 Change
Standard	.30	.30*
Contexts Test	.54	.24*
Levels Test	.61	.05*
Standard	.30	.30*
Levels Test	.55	.25*
Contexts Test	.61	.06*
Levels Test	.51	.51*
Contexts Test	.58	.07*
Standard	.61	.03*
Contexts Test	.49	.49*
Levels Test	.58	.09*
Standard	.61	.03*

*$p < .05$.